SP-2057

SP *a professional paper*

The Conceptual Foundations of Information Systems

by

H. Borko

SYSTEM DEVELOPMENT CORPORATION
2500 COLORADO AVE.
SANTA MONICA
CALIFORNIA

Paper to be read at the symposium: The Foundations of Access to Knowledge to be held at Syracuse University, July 28-30, 1965.

TABLE OF CONTENTS

		Page
I.	Information Systems and Sciences	5
II.	The Structure of an Information System	6
III.	Basic Concepts of the Automated Information System	9
IV.	System Implications Concerning Automation--Concepts #1 and #2	11
V.	System Implications Concerning the User--Concept #3	14
VI.	System Implications Concerning Automated Language Analysis--Concept #4	15
VII.	System Implications Concerning Automated Indexing--Concept #5	18
VIII.	System Implications Concerning Automated Classification--Concept #6	20
IX.	System Implications Concerning Document Storage--Concept #7	23
X.	Restatement of the Concepts on which Information Systems are Based	25
XI.	A Look at the Information System of the Future	26

TABLE OF FIGURES

Figure		Page
1.	An Information Storage and Retrieval System	7
2.	Word Lists as Derived by Statistical Analysis of Text	17
3.	An Example of An Association Map Index	21
4.	Remote Inquiry Station	31
5.	First-Level Display	32
6.	Light Pen Action	33
7.	Second-Level Display	34

(page 4 blank)

ABSTRACT

Information systems consist of collections of recorded information, custodians who organize and maintain the collections, retrieval procedures and users. The conceptual foundations for these systems are derived from mathematics, engineering, behavioral science and the many other disciplines which together make up information science. The concepts are the theoretical formulations or principles concerning methods of storing, indexing, and retrieving information which are used in the design of information storage and retrieval systems. Seven concepts are enunciated. These deal with the need, equipment user responsiveness, language processing, indexing, classification and storage. The system design implications of each concept are discussed separately and then organized together to form an information storage and retrieval system of the future called BOLD.

INFORMATION SYSTEMS AND SCIENCE

An analysis of information systems must begin with an understanding of information science, for the theories of information science form the conceptual foundations of information systems.

Information science is the discipline--the theoretical discipline--concerned with the applications of mathematics, system design, and other information processing concepts. It is an interdisciplinary science, involving the efforts and skills of librarians, logicians, linguistics, engineers, mathematicians, and behavioral scientists. The application of information science results in an information system. The role of information science is to explicate the conceptual and methodological foundations on which existing information systems rest and to develop new concepts on which improved systems can be based. Information technology, as contrasted with information science, involves the application of tools and techniques to the operational problems of information systems.

For purposes of our discussion, let us agree that an information system consists of a collection of recorded information, custodians who organize and maintain the collection, a retrieval procedure, and the users who refer to the information to satisfy a variety of needs. As this definition implies, there is a great deal of similarity between a library and an information system. Indeed, there must be, for a library is a specific type of information system with a collection of documents, a characteristic method of organizing and maintaining the collection, and a designated set of users. In contrast, an information system refers to a more generalized complex of functions.

The exponential increase in the amount of scientific and technical documentation, and in the need for people to be aware of advances in science and technology, is providing the impetus for growth in information science. Concurrently, the availability of electronic data-processing equipment is providing new and more powerful tools to cope with the increase in documentation. The task of information science is to re-examine existing methods of acquiring, storing, indexing, and retrieving information in the light of advanced technology and to derive new concepts and principles that can be used in the design of more efficient information storage and retrieval systems.

THE STRUCTURE OF AN INFORMATION SYSTEM

The general structure of an information system can be illustrated by a simplified block diagram.[1] Printed data are received and analyzed to determine which index terms are to be associated with the documents as well as the subject categories into which the documents are to be classified. The analysis may also involve the translation of the document from one language to another and the preparation of an abstract.

A significant feature of the diagram is the branching that occurs after the completion of analysis. This branching allows for more than one method of organizing the file. The documents themselves may, for example, be physically arranged on the shelves by subject classification or by accession number. Since only one physical organization is possible without expensive duplication, alternative paths of access to the file will be by reference to the

[1] This diagram is a modified version of one which appeared originally in J. Becker and R. M. Hayes, Information Storage and Retrieval, [1] p. 70.

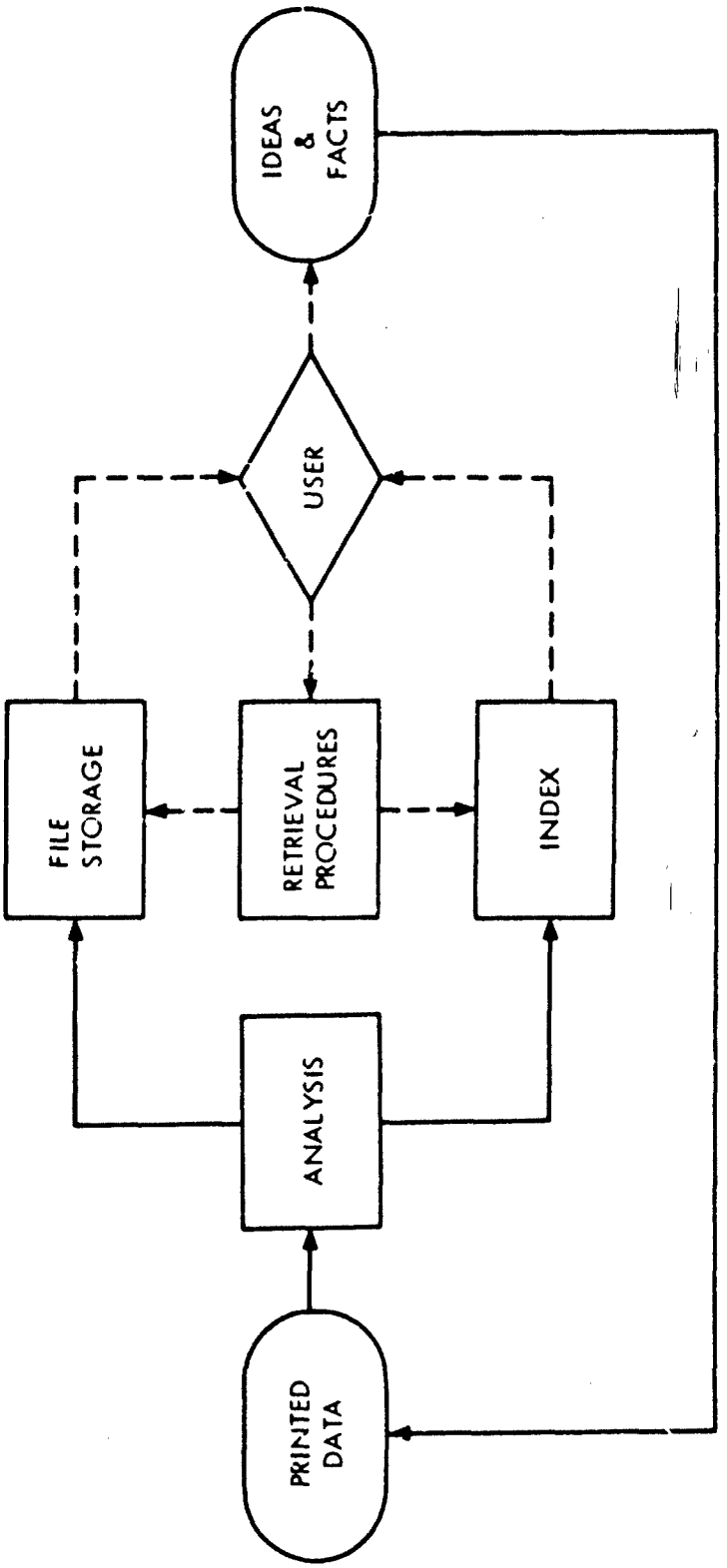

FIGURE 1. AN INFORMATION STORAGE AND RETRIEVAL SYSTEM

document, e.g., by card catalog rather than by access to the documents themselves. The physical arrangement of the documents is represented in Figure 1 as the storage file and the reference organization is represented as the index.

The user requests information. To obtain it, he interacts with the index and/or file. To do so he must convert his request for information into a well-defined search question and retrieval procedure to which the system can respond. In an open-stack library, the search request may be implicit and the retrieval procedure simple: He searches in the card catalog for either a specific item or a class of items, and proceeds to the shelf location of the document for which he is looking. Or, the librarian performs the retrieving task in response to an explicit request from the user.

When the user is satisfied with the documents he has been given, he reads them, to obtain new ideas and facts. Presumably, he will now become a generator of new printed data, which will in turn find its way into the system, and the cycle is complete.

The differently shaped boxes in the diagram also have meaning. The beginning and end points of the system, namely, printed data and ideas and facts, are encased in ovals, to indicate that these are areas of concern to specific disciplines, and are often neglected in all but the broadest approaches to information science. For example, within the heading of printed data, one would have to consider the principle underlying the acquisition policy. For the moment, let's relegate this to the librarian. Similarly, we will let the psychologist worry about how ideas are derived from documents.

The user represents a unique set of problems, and this uniqueness is recognized by the diamond-shaped box. The information system is designed to satisfy the needs of the user and these needs differ from person to person and from problem to problem. Furthermore, the user is the criterion by which the information system is evaluated. If the user does not get the service he can expect, the system is not fulfilling its function. We will touch upon the determination of user requirements during the discussion of methodologies of system design. At any rate, it is obvious that the user is an integral part of the information system and that he requires special treatment.

The other parts of the system diagram, namely, analysis, file storage, index and retrieval procedures, are all in rectangles. These, together with user requirements, constitute the main portions of the information system. Let us begin our discussion of an automated information system by examining the basic concepts on which the system rests.[1]

BASIC CONCEPTS OF THE AUTOMATED INFORMATION SYSTEM

Concept #1, Need for an Information System: If we are to cope with the expanding volume of information, the present methods of operation must be changed and modernized. Computer processing of natural-language text--called automated language processing--will play a central role in the information system of the future.

[1] The formulation of these concepts was strongly influenced by Merrill Flood's paper "The Systems Approach to Library Planning" [4]

Concept #2, Equipment: While it is assumed that the general-purpose digital computer will be used in the modern information system, design considerations should not be limited to this piece of equipment. Almost any device that can be conceptualized can be developed, if time and cost are not important. This gives the system designer flexibility in specifying his requirements, but eventually the system will have to be evaluated and justified economically.

Concept #3, User Responsiveness: An information system is designed to satisfy the needs of the user, and therefore knowledge of user needs is a prerequisite. Since the user has varying needs for information, the system should be flexible and capable of many modes of operation, including browsing through the accumulated store, searching for specific information, and locating single documents. Ideally the system should also play an active role and notify users that items of possible interest have been received and are available for distribution.

Concept #4, Automated Language Processing: Computers are symbol manipulators and can be programmed to process words as well as numbers, and thus to analyze language. At the simplest level, the computer can be programmed to count word-tokens and provide frequency distributions of word-types, e.g., the number of times the word "program" is used in this paragraph. On a more sophisticated level, existing programs for syntactical analysis of text enable one to deal with relationships among words, and some initial attempts have been made to process semantic information.

Concept #5, Indexing: Indexing provides a short descriptive tag of the information contained in the document. There is, as yet, no theory that tells one how to code an item in a collection so as to maximize the likelihood of successful retrieval over expected queries. All existing indexing codes are approximations of the total information content. For efficient operation, a capability must be provided for changing the index tags when necessary.

Concept #6, Subject Classification: There is no economical way to scan the contents of a large collection in its entirety, in response to every query. The collection must be searched selectively; the purpose of subject classification is to divide the store into reasonable sections to facilitate search.

Concept #7, Storage: The physical documents are to be stored in permanent locations compactly and economically. Shelving will probably be by accession number, modified by special space and handling requirements as in the case of oversized books, engineering drawings, etc. An auxiliary storage file will contain a complete microfilm collection of all available documents, for search and for hard-copy reproduction.

SYSTEM IMPLICATIONS CONCERNING AUTOMATION--CONCEPTS #1 and #2

Having defined some of the basic concepts and outlined a general information processing system, we can now examine the system design implications of these concepts. Concept #1 states that we must modernize and automate information handling procedures if we are to cope with the expanding volume of information. Concept #2 states that the general purpose digital computer can be used to process language data, and that future models may be even more appropriate for

the task. Certainly our thinking should not be constrained by equipment limitations. The automated information system must be able to satisfy the many different needs that a user has for information, and to do this more efficiently than existing systems. What are these needs? A number of studies have analyzed the information requirements of users. The different types of user requests and their implications for system design can be summarized as follows:

Requests for a Specific Document

The most frequent request received at an information center is for a specific document. Libraries are geared to handle this kind of request, but occasionally have difficulty in supplying the document, which may be on loan or misshelved. In the automated system the user obtains the shelf location or document number from an index file and receives a copy of the document from the intact microfilm file.

Requests for Information by Subject

Another rather common request is for information on a particular topic, such as documents dealing with system design concepts. Most of the research in information storage and retrieval has been concerned with improving the system response to this form of request, and a great deal of progress has been made. Generally, the user wants relevant documents, but not necessarily all relevant documents, and he wants the material in a reasonable period of time. Most information systems have no difficulty in responding to this request.

Requests for an Exhaustive Search

Some research problems require that the investigator make a systematic and thorough search through all of the literature, to determine whether the information he needs has been reported elsewhere. Sometimes the user hopes that he will not find the data, for instance, in patent searching where the inventor wants to make sure that no one else has patented his idea. There is no logical way of insuring that all relevant documents will be retrieved, short of screening the complete file of documents. Different search strategies and retrieval procedures are being studied, and need further study and comparison, to maximize retrievability of all and only the relevant documents.

Browsing

The user searches for items that are interesting, original, or stimulating. No one can find these for him; he must be able to browse through the data himself. In a library, he wanders among the shelves picking up documents that strike his fancy. An automated information system must provide similar capabilities. How this can be accomplished is still a matter of concern to the researcher, and no final answers are available.

Current Awareness

Finally, users request information to keep abreast of developments in a given field, and there are now several journals (Chemical Titles, for example) devoted to current awareness in specialized areas. Also various selective dissemination procedures attempt to provide information to the specialist before he recognizes his own needs. These services should be integrated into information system.

SYSTEM IMPLICATIONS CONCERNING THE USER--CONCEPT #3

Recognizing that there are different categories of information and that different people, or the same person at different times, will have changing needs for information, the system designer must know and be responsive to the needs of the users.

For information on procedures for gathering data on information use, we turn to the social scientist who is concerned with the accessment of human behavior in a variety of situations. There are four basic techniques of observing behavior in a controlled fashion so that the results can be verified objectively. These are (1) field study, (2) case study, (3) survey, and (4) experimentation. Each of these can be applied to the study of user requirements.

In the field study approach, the observer stations himself at the information center or reference library and studies the behavior pattern of the user as he searches for information. Does he use the card catalog? Does he ask the librarian for help? This is a simple and effective method of gathering data, but it is time consuming and difficult to obtain a statistically adequate sample of all the categories under investigation.

To get a more representative sample of users, as well as a more longitudinal study of information use, the case study method is employed. Here selected individuals are asked to keep records of their use of information, but this too is a slow process.

If ones needs to speed up the acquisition of data on user requirements for large numbers of individuals, the survey method is used. The survey, which may consist of mailed questionnaires or of oral interviews, can provide detailed information.

Finally, one can set up controlled experiments to compare and evaluate different situations, as in the case of the Cleverdon studies of different indexing systems [2].

These are the classical methods of scientific observation and must be part of the methodological tools of the information system designer.

SYSTEM IMPLICATIONS CONCERNING AUTOMATED LANGUAGE ANALYSIS--CONCEPT #4

Ultimately, an information system is concerned with the analysis of ideas expressed in documents. Because of the expanding volume of information, an automated system will be required (Concept #1). In this system, the computer will be used to process the natural language of the text and, thus, play a central role in the analysis of documents and in the information system of the future (Concepts #3 and #4). For guidance on the principles and methods of automated language analysis, we turn to the linguist and to two relatively recent offshoots of linguistic science--computational linguistics and psycholinguistics.

Ideally, we would like the computer to scan the full text of a document and analyze its contents, to provide indexing terms, subject headings, a classification, an abstract, and even a translation into another language.

What techniques are available, and how can they be utilized in the information system?

In the present manual mode, the analysis of a document and the writing of translations, abstracts, index terms, etc., are complex processes carried out by skilled humans who read the original document and who understand the ideas contained therein. The computer is a machine and not a "giant brain"; it can manipulate words but not ideas per se, and is dependent upon the knowledge it has been provided.

Recognizing these limitations of our tool, the questions still remain, what linguistic principles are available and what techniques can be developed to index a document, to write an abstract or translation, automatically.

The basic concept in language analysis is that meaning is carried in the words and in the arrangement of the words used in the document. The techniques of language analysis stem from this concept. Statistical techniques count individual words and word pairs; syntactical techniques deal with the arrangement of words in sentences.

To illustrate the power of statistical analysis, I prepared, by computer program, a list of content words derived from two articles in a child's encyclopedia [see Figure 2]. Even though these words are presented in list form, one experiences little difficulty in determining the subject matter of the articles. Yet, without the arrangement of words or the sentence syntax, one loses a great deal of information. To take an overworked

DRUMS	DANCING
African	Ballet
Band	Ballroom
Bass	Dance
Dancer	Folk
Drum	Music
Head	Opera
Kettledrum	People
Leather	Religion
Messages	Social
Metal	Waltz
Orchestra	
Player	
Snare	
Sound	
Tambourine	
381 words	278 words
202 word types	146 word types

FIGURE 2. WORD LISTS AS DERIVED BY STATISTICAL ANALYSIS OF TEXT

example, the two sentences, DOG BITES MAN and MAN BITES DOG contain identical words and yet the meaning is different. One needs to know the words and the arrangement of the words, or the syntax, of the sentence. Computers can be programmed to perform both statistical and syntactical analysis of language.

SYSTEM IMPLICATIONS CONCERNING AUTOMATED INDEXING--CONCEPT #5

As stated in Concept #5, the purpose of indexing is to provide a short descriptive tag of the information contained in a document. Regardless of whether these index tags are supplied by skilled human indexers or as a result of some machinations by a computer program, they will not be complete representations of the document content because they are highly abbreviated. Nor is there any way of knowing in advance that one set of index terms will be better retrieval tags than another set. There exists no theory for optimizing the indexing function. This is not to state that there are no differences between indexes or between human and mechanical indexing. There are differences, and there may be advantages and disadvantages to each index, but one form is not automatically better than another.

The basic concept of automated indexing is that the documents should generate their own index terms. One method of accomplishing this is by means of a statistical analysis of the words in the document. This procedure was suggested by H. P. Luhn [6] in 1957 and it remains the fundamental technique.

Figure 2 illustrates the results of statistical indexing. One obtains a list of words that obviously have something to do with the content of the document. I will not go into the question of whether or not an unedited list of words

should be used as the index to the document, because this depends on the total system and the manner in which the words are to be used, but at the very minimum, I believe we will all agree that such a list can be useful as an aid in assigning index terms.

A basic concept of automated indexing is that uncommon words that appear frequently in a document are suitable candidates for use as index terms. This is a necessary but not a sufficient rule for automatic indexing. For example, the rule does not specify the meaning of the word "frequently"; how many times must a word be used before it is accepted as an index term? Obviously, additional rules need to be formulated and tested.

One of the difficulties encountered in using word-count techniques for producing an automatic index is that they invariably result in large numbers of index terms. However, these lists could be pruned by human post-editing to make the number more reasonable.

Of even greater concern is the fact that an index term is not meant to be a word but rather a concept, so the words ballet, waltz, ballroom all refer to styles of dancing. Can the computer combine these words and phrases automatically to form a concept that will be used in indexing the document? Possibly. Some researchers are attempting to form concepts, and some methods have been suggested. These methods of grouping words include the use of thesauri, associative indexing, factor analysis, clumping procedures, etc.

Allow me to point out, once again, that there is no <u>a priori</u> reason for insisting that the machine-produced index imitate the traditional human

prepared index. The computer can prepare other forms of useful indexes that the human cannot do as well. The permutation index of titles is one such example. Luhn called this the KWIC index, and it is now fairly common. Another approach is the citation index, which lists all articles that have referred to a particular paper as well as the articles cited by that paper. This form of index is particularly useful in research work.

An even more radical form of index is the association map suggested by Doyle [3] (Figure 3). The standard index is an alphabetically arranged list of index terms, but the user is not usually interested in finding words that start with the same letter of the alphabet. He is trying to associate topics. With this in mind, Doyle devised a two-dimensional map-like index, which pictures the association among the terms by use of connecting arrows.

Undoubtedly, still other forms of automatically and semi-automatically produced indexes will be developed and their use will become more acceptable and widespread.

SYSTEM IMPLICATION CONCERNING AUTOMATED CLASSIFICATION - CONCEPT #6

Closely allied to automated indexing is automated classification.

Classification provides an orderly arrangement of books, documents, or other material, so as to facilitate retrieval. The most important aim of information science is to provide faster and more precise access to the existing store of information. Since the total collection is large, the search must be done selectively. Classification is necessary to divide the information into reasonable portions, and to arrange the contents for efficient access.

ASSOCIATION MAP

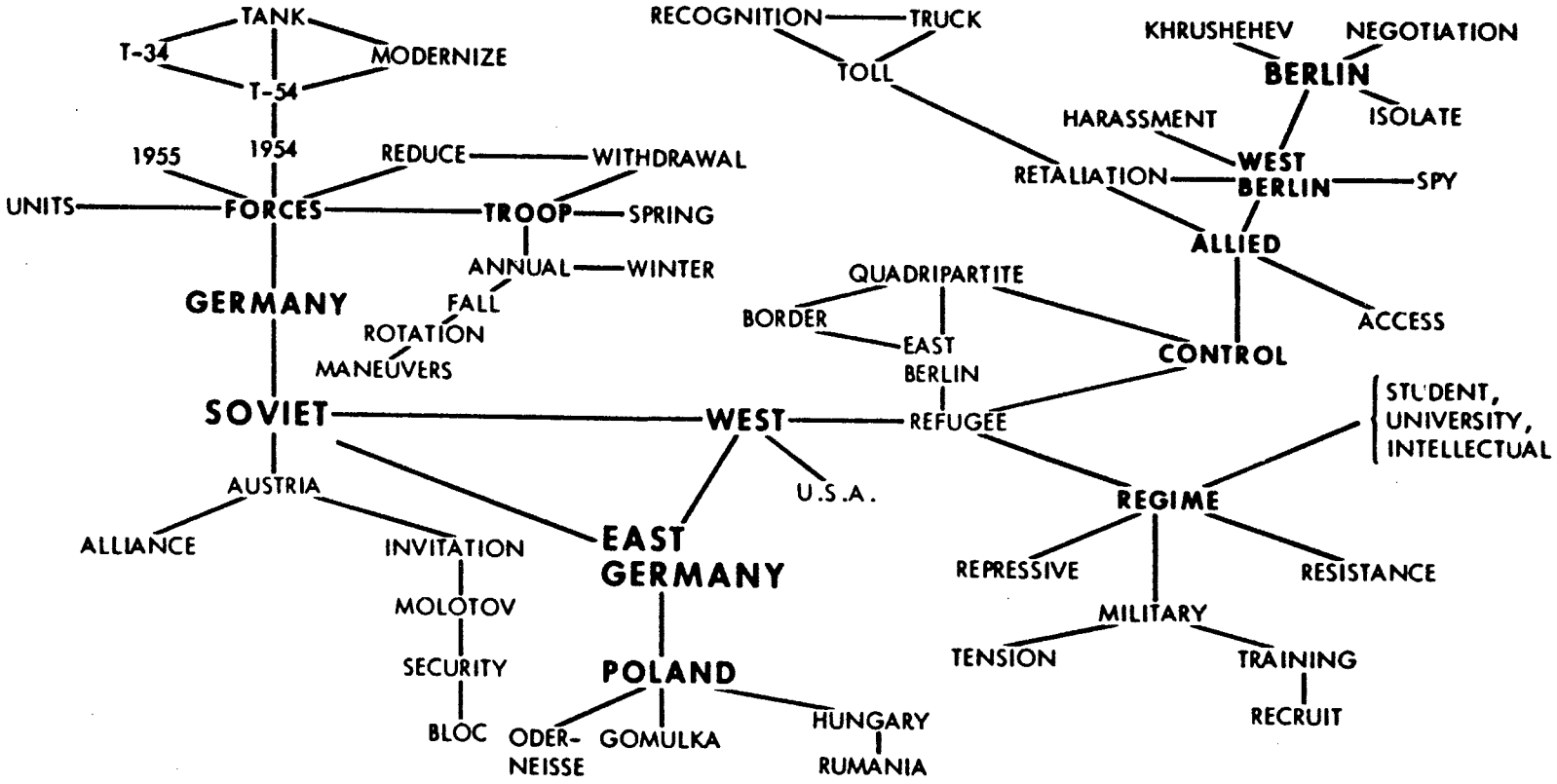

FIGURE 3: AN EXAMPLE OF AN ASSOCIATION MAP INDEX

If we think in terms of library classification systems, we realize that the Dewey Decimal System, the Library of Congress System, and the others, are all schemes for arranging books by subject so that the reader can find what he wants efficiently. Even in a fully automated system, some form of classification is necessary to keep search time within reasonable bounds.

Research in automated classification consists of two complementary projects: those dealing with automatic methods of deriving a classification system and those concerned with automatic methods of classifying documents into existing categories. The first project is based on the concept that classification systems can be derived empirically, and studies methods of mathemathically analyzing a document collection to determine the number and description of classes to be established. The second project is concerned with the use of automatic methods of classifying documents and is based on the concept that the similarity among documents can be measured and that procedures can be developed for computing the probability of a document belonging to a given category. In general, documents are designated as belonging to the same class if they are more similar to each other, as determined by the words they share in common, than they are to documents in any other class.

A number of investigators, both in the United States and in Europe, are actively engaged in research in these project areas. Laboratory models for mathematically deriving classification systems and for categorizing documents already exist. Progress is being made toward the development of a fully automated document classification system. The information system

designer must be able to integrate these procedures into his system to provide increased user responsiveness.

SYSTEM IMPLICATIONS CONCERNING DOCUMENT STORAGE--CONCEPT #7

Shelving documents by subject facilitates browsing through the open stacks and makes retrieval more efficient. However, there are also disadvantages to this form of storage. One must allow space on each shelf for storing new additions to the library. Also a change in the classification system, e.g., from Dewey to Library of Congress, requires a complete reshelving and renumbering of all documents. This is expensive and time-consuming.

In an automated information center, which is basically a closed-stack library, we have assumed that the physical documents will be stored by accession number in permanent locations. We have further assumed that a complete microfilm file of the collection will be available for search and hard-copy reproduction. Certainly, a microfilm file will be more compact and economical to store than the collection of original documents. The question is whether such a file can be searched efficiently. Efficient search strategy requires that documents be grouped by subject, but economical storage uses unclassified compact shelving. One way to resolve these contradictory requirements is to make a microfilm image of the document and to use the microfilm in the search and retrieval procedures. The microfilm copy can be manipulated and rearranged easily. It is this easy-access file that can be organized by subject category to facilitate browsing and retrieval while the original documents are arranged by accession number.

Unfortunately little research has been done on methods of organizing and structuring large files. As a result, the available concepts are primitive. One principle is clear: our needs for information are changing; therefore, our file organization must be capable of changing. The division of all knowledge into ten major categories as suggested by Dewey may no longer be adequate. The various empirical methods of devising classification systems suggest other procedures for organizing files. These have the advantage of being able to quickly reorganize the entire file to make it more responsive to the needs of the user. Flexibility is all well and good, but it still leaves unanswered the question of what is an efficient file structure.

The typical file structure is based upon a hierarchical subject classification, a structure that is familiar and allows for the use of a simple search strategy. However, Hayes [5], suggests a different concept based upon the activity of the file items. His notion, stated somewhat oversimply, is to place the items most likely to be used in the most accessible positions of the file. This method requires that requests for items be recorded and that the file be updated systematically according to use patterns. Assuming that some reference to the file is in a machine-usable language, the updating can be done easily. The activity method of file organization has an added appeal, because it provides the data for systematically purging the active file of rarely used items.

RESTATEMENT OF THE CONCEPTS ON WHICH INFORMATION SYSTEMS ARE BASED

We have now discussed the various parts of an information system. Let us stop for a moment and review the concepts on which the system is based.

Need for an Automated Information System

The very first concept states that there is a need for an automated information system. Present methods of gathering, storing, and retrieving information will soon prove inadequate to cope with the ever increasing volume of published literature. New techniques are needed and these innovations involve the use of computers to process documents in their natural language. Should present-day computers not be adequate for the task, it is assumed that new pieces of equipment, with greater capabilities, could be designed and built.

The concept on which automated language analysis rests is that ideas are communicated by words and the arrangement of words as written or spoken. Stemming from this fundamental tenet are the related design concepts that the meaning of the document, or at least the major ideas that the author is trying to communicate, can be determined by a mechanical analysis of the words and the arrangement of the words used. By means of statistical and syntactical techniques of language analysis it is possible to automatically index, classify, abstract, and translate natural language. While it is recognized that existing techniques fall far short of achieving these goals, these concepts form the basis for concluding that automated language analysis is logically possible.

User Responsiveness

The information system must be versatile and capable of satisfying the many different needs of the user. The design concepts concerning the user deal with the methodological problems of determining user requirements for information. These concepts were borrowed from the social scientist who has developed procedures for observing and measuring complex behavior under controlled conditions. User needs can be determined objectively by field study, case study, survey, or a combination of these methods.

Automated Indexing

Indexing consists of selecting a series of words and phrases that will identify the information contained in the document. As stated previously, there is no theoretical basis for selecting index terms so as to guarantee that a document so indexed will be retrieved in all cases, and in only those cases, when it will be relevant to a user's request. This is another way of stating the obvious fact that the few index terms can not serve as a representation of the total document content. It is also pointed out that there is no theoretical basis by which one can optimize the selection of index terms--that there is no one best set of terms. While this may be discouraging for those who seek perfection, it provides a license for those who would explore various ways of preparing an index.

The design concept of automatic indexing is that index terms can be selected by an analysis of the words used in the documents. This analysis may be accomplished by a simple statistical counting of the words and selecting

frequently used non-common words. Any additional processing that needs to be done can be accomplished by the use of a thesaurus, associative indexing and clustering techniques and--at least for the present--by human post-editing.

New and different types of indexes can and are being produced by computer aids. These include Uniterms, Key-Word-In-Context, Association Maps, and Citation Indexing.

Automated Classification

Classification is used to arrange data, or documents, in an orderly manner to facilitate retrieval. Large collections must be searched selectively and so, must be divided into sections for efficient use. Manual classification is based on logical and reasonable systems for dividing knowledge into categories. Automatically derived classification systems are based on the use of mathematical techniques for grouping documents on the basis of similarity of word usage. The techniques used include clump theory, factor analysis, regression analysis, latent structure analysis, etc. Since the basis of manual and automatic classification are different, the resulting systems are also different, although both may be useful.

There are two facets to the research in automated classification. One is concerned with the mathematical derivation of classification categories; the other studies the use of automated techniques for determining the category into which the document should be placed. The concept involved is that documents belonging to the same class have more content words in common with

each other than they have with documents in other categories. By using factor scores or Baysian prediction equations, one can predict the probability of a document belonging to a category.

Since the automated classification system has been derived by analysis of a given document collection, it is a relatively simple matter to reorganize and reclassify the collection when a significant number of new documents have been added. Not only will it be possible to determine, by mathematical computation, whether new categories need to be started or old ones combined, but it will also be possible to regroup the complete microfilm file of documents into the new categories.

Document Storage

The physical documents in the collection should be stored compactly to make efficient use of the available shelf space. When searching or browsing for information, the requester will make use of the microfilm file, organized by subject and activity level. If hard copies are desired they will be reproduced from the microfilm.

A LOOK AT THE INFORMATION SYSTEM OF THE FUTURE

Let us take another look at the schematic diagram of an information storage and retrieval system and see how the assumptions and concepts that we have discussed can be fitted together into an automated information storage and retrieval system.

The printed item--a document--enters the system. Actually, it arrives in two different formats, the traditional printed version and as a reel of magnetic tape, since, in all probability, the document will be typeset automatically by computer. The printed copy is examined for acceptance, microfilmed, and placed, by accession number, in storage. The magnetic tape version is run through a computer program where it is indexed in depth, abstracted automatically unless an author-prepared abstract exists, and assigned to one or more subject classification categories. While these analyses can be done automatically, the information specialist will be able to review the results, add or delete terms, and make whatever other decisions are deemed desirable. The microfilm copy of the document, complete with its index terms and abstract is placed at the head of the other documents in the appropriate subject file. It is placed at the head of the file, since our initial hypothesis is that new material will have a higher activity rate than will older material.

Each week, or at some other regular period, the index terms for the new acquisitions are matched against a user profile. Where matches occur, notices of the new acquisitions are prepared automatically and selectively distributed to interested individuals. These users may then order microfilm or photocopies of the document as they wish.

Major users or organizations will have remote inquiry stations, consisting of a teletypewriter and a display console. To request documents, the user dials the large time-shared computer at the information center. Prototypes

of parts of such an automated information center already exist at MIT, as part of Project MAC (Multi-Access Computer); at Harvard in SMART (Salton's Magical Automatic Retriever of Texts); and at System Development Corporation in the BOLD (Bibliographic On-Line Display) System.

Figure [4] is a picture of the remote console station as used in the BOLD System. Imagine that the information center is situated in California and that you, at some remote location, are sitting at this inquiry station. You want some information on fungus diseases but you are not really sure that the information center has any documents on this topic. You might start your search by asking the computer what categories of documents it has on file. You do this by typing your question on the keyboard. Your response would appear on the display tube as illustrated [Figure 5].

Note that this file is organized according to the Library of Congress classification system. In later versions, the data base or document collection may be organized into mathematically derived categories or Universal Decimal classes. At any rate you note that you are to select the category in which your information is most likely to be found. Quite clearly you are interested in category R for medicine.

Using the light pen as illustrated [Figure 6], you communicate your selection to the computer by flashing the first letter of the selected category.

This results in a new display [Figure 7].

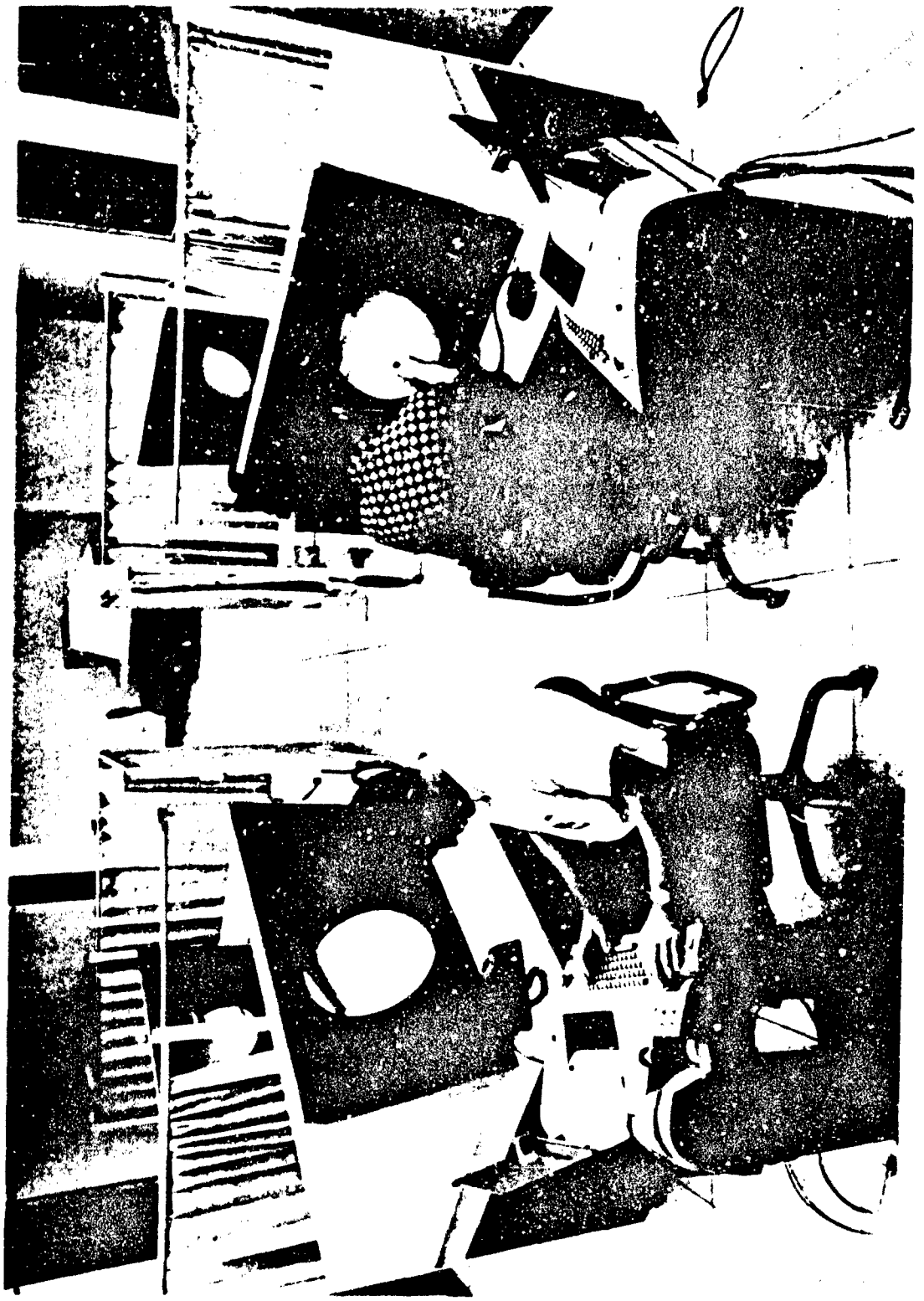

FIGURE 4. REMOTE INQUIRY STATION

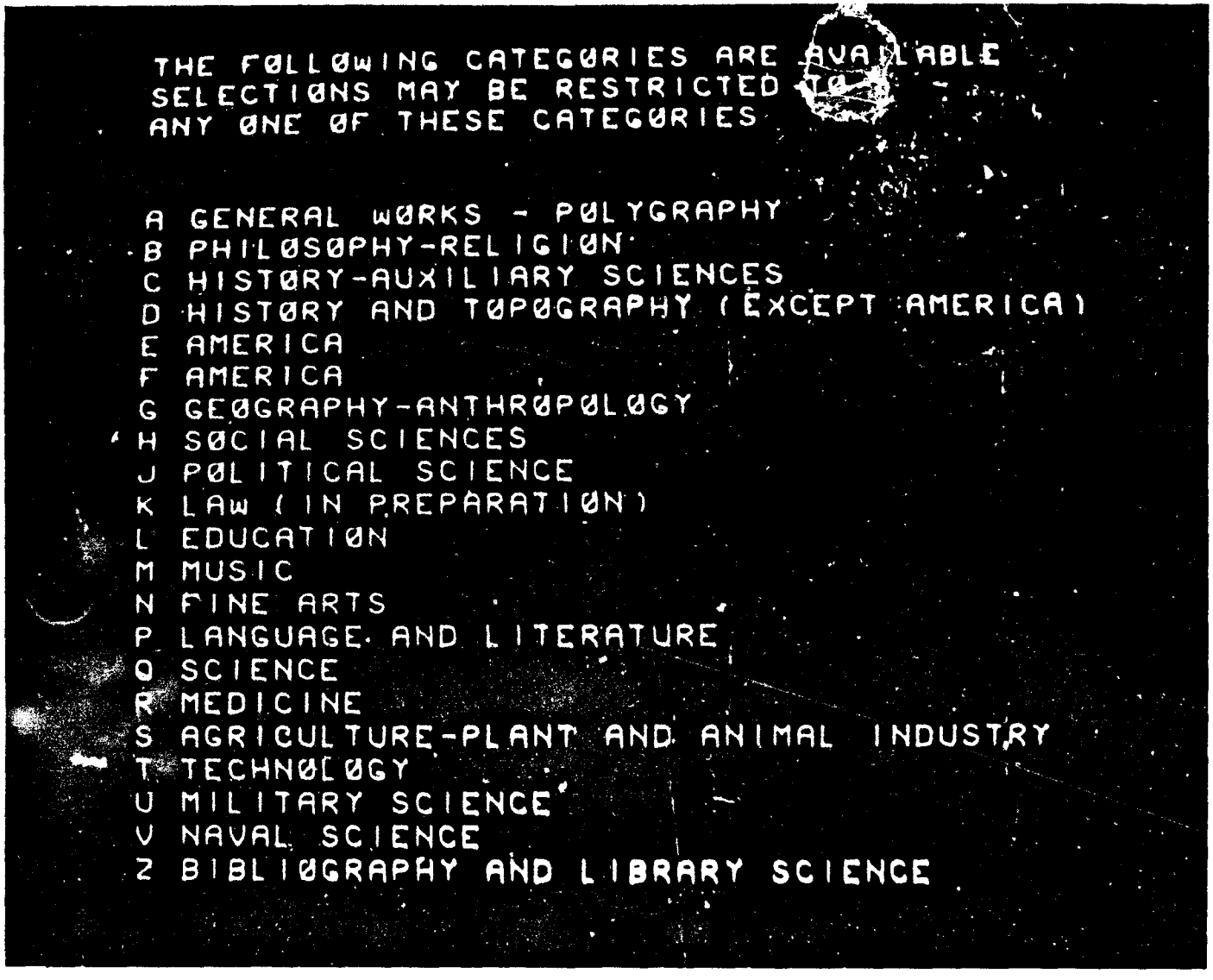

FIGURE 5. FIRST-LEVEL DISPLAY

FIGURE 6. LIGHT PEN ACTION

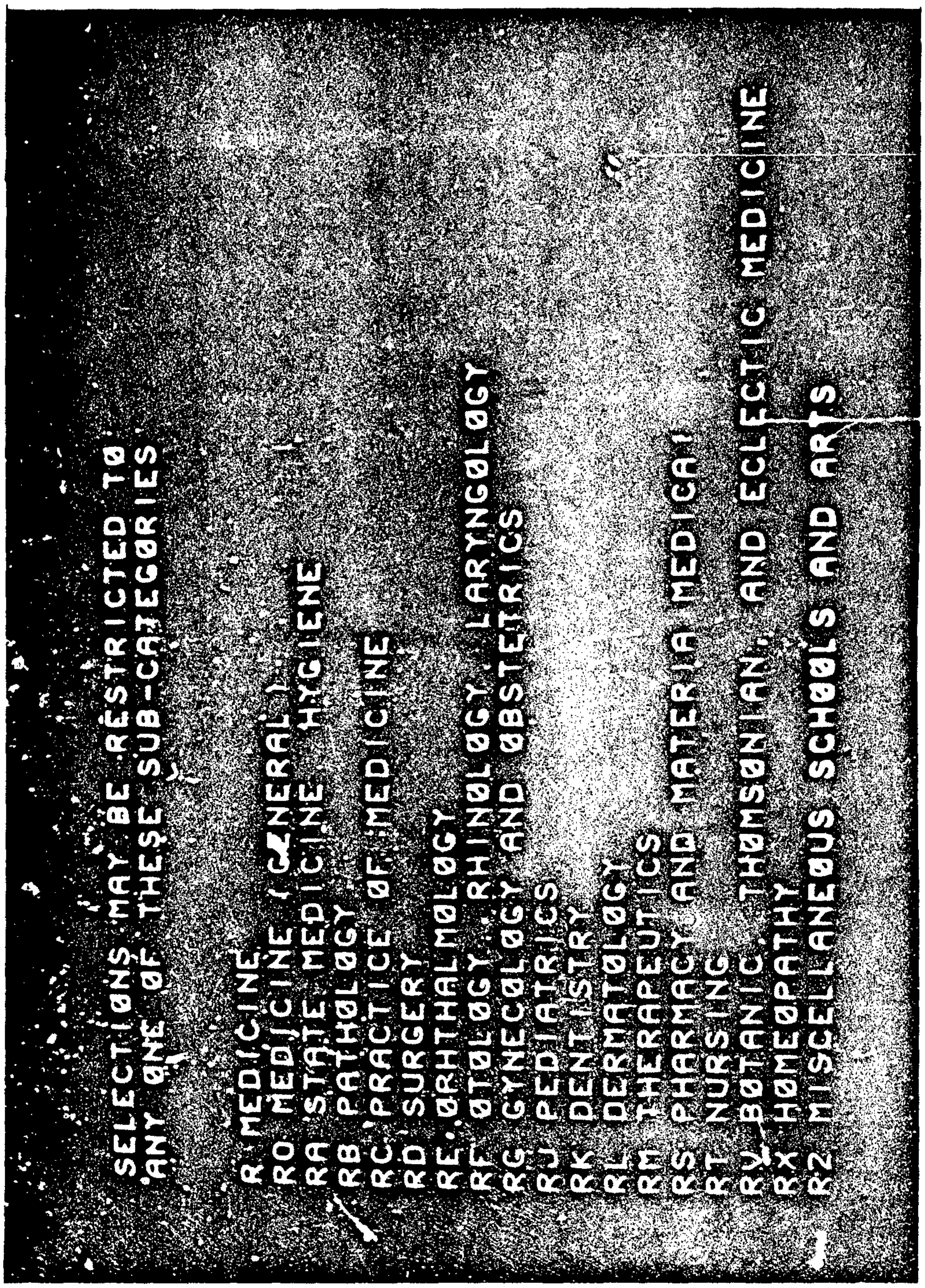

FIGURE 7. SECOND-LEVEL DISPLAY

Now you see the listed sub-categories, and again you use the light pen to make your selection. In this manner you narrow your field and reduce the number of documents to be searched. You may then ask to see the documents in the sub-category that you have selected. On the scope you will see the author and title of the books which have been classified into that category. You may now browse through the file and make your selection. On request, you will be given the appropriate call number and you may ask for a microfilm copy. The microfilm image could be transmitted to you on the display console or you can request that the film, or a photocopy, be mailed to you.

And that is not all!

Frequently in our search for information, we are not interested in browsing, but we look for documents dealing with a specific topic. We ask for these by listing a string of index terms. The computer will search its file and report how many documents contain all of the desired terms and, thus, meet the specified criteria. Before requesting the documents you have the option of adding, deleting, or changing terms and thus decreasing or increasing the number of documents that you will receive. When you are satisfied with your request, you may ask to see the author and title of the selected documents. You may now eliminate those documents that you do not wish to read either because you are already familiar with them or because they seem to be

tangential to your interest. If you need additional information to help you decide, you may ask to see the abstract. By means of this man-machine dialogue, you are helped to select only relevant documents.

Although the information center which I have just described does not yet exist, its prototype is now being programmed in accordance with the concepts and design principles we've just described. The future is not too many years away.

BIBLIOGRAPHY

1. Becker, J. and Hayes, R. M. Information Storage and Retrieval. New York: John Wiley & Sons, 1963.

2. Cleverdon, C. W. Report on the testing and analysis of an investigation into the efficiency of indexing systems. Cranfield, England, 1962.

3. Doyle, L. B. Semantic road maps for literature searchers. J. of the ACM, Vol. 8, No. 4, 1961, 553-78.

4. Flood, M. M. The systems approach to library planning. The Library Quarterly, Vol. 34, No. 4, 1964, 326-28.

5. Hayes, R. M. The development of a methodology for system design and its role in library education. The Library Quarterly, Vol. 34, No. 4, 1964, 339-51.

6. Luhn, H. P. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, Vol. 1, 1957, 309-17.